

NTSEBench: Cognitive Reasoning Benchmark for Vision Language Models

Pranshu Pandya[†], Vatsal Gupta[†],
Agney S Talwarr , Tushar Kataria , Dan Roth, Vivek Gupta^{*}

[†]Equal Contribution, ^{*}Corresponding Author

Indian Institute of Technology-Guwahati,
University of Utah, University of Pennsylvania, Arizona State University



Indian Institute
of Technology,
Guwahati



MOTIVATION

- The Need for **Cognitive Reasoning** in the AI Landscape
- Advancing AI Toward **Human-Like Problem Solving**
- **NTSE Exam-Based Dataset**
 - **The National Talent Search Examination (NTSE)**: Taken by 1 million students annually in India.
 - **Evaluates**: Critical thinking, analytical reasoning, general knowledge, and mathematical aptitude.
- **AI Training**: Use NTSE-style questions to test Multimodal models on deep cognitive reasoning, drawing a parallel to human problem-solving skills.

ability strategies
patterns learning propriety
NTSEBench baselines
extensive
curatedquestions
training multiple decipher
quickly assess nationwide
general choice
analogies tasks
drawn state
source multimodal
language four new critical
different comparison
reasoning NTSE
evaluate deeper
LLMs open India
categorized distinct
struggle puzzles types
excel modalities
human facilitate
amounts establish
images series problem
dataset beyond thinking
textual understanding
common solving examination

INTRODUCTION

- NTSEBench is a **novel benchmark** designed to evaluate cognitive reasoning in vision–language models.
 - NTSEBench aims to tackle **gaps in current benchmarks**
 - The dataset targets **advanced skills**—such as pattern recognition, logical deduction, and spatial reasoning—that go beyond rote memorization.

ability patterns NTSEBench skills learning propriety extensive data baselines

curated questions large decipher nationwide

quickly assess choice general choice

analogy tasks state general choice

drawn source state general choice

language multimodal new critical comparison

different critical comparison

reasoning intelligence NTSE

evaluate deeper

cognitive challenges categorized

struggle facilitate

LLMs open India puzzles distinct text types

models

excel human modalities

amounts series problem

images establish thinking

dataset

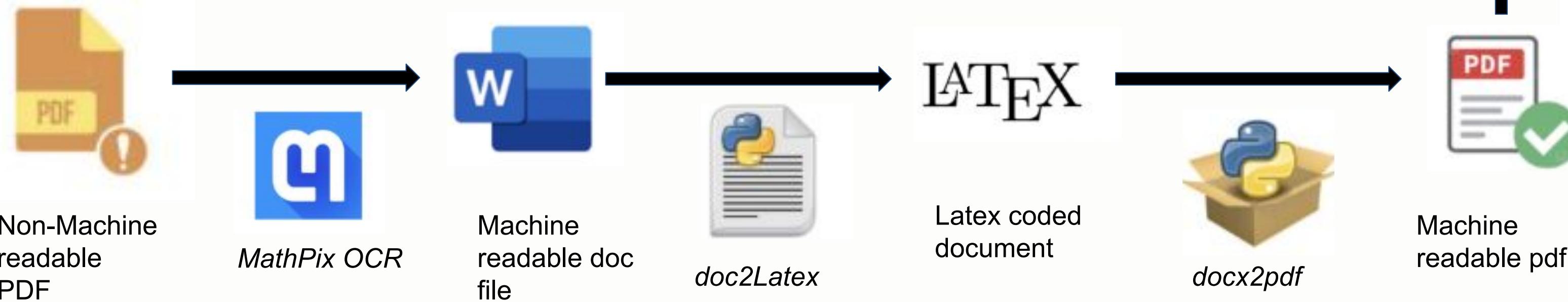
textual understanding

beyond common solving examination

common spatially

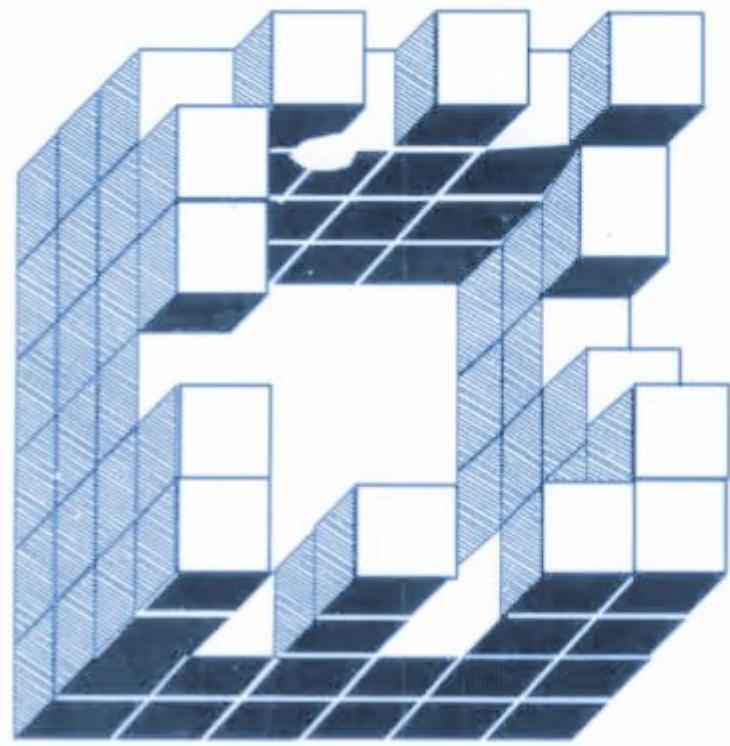
DATASET CURATION - PIPELINE

- **Manual curation & extraction pipeline** to convert non-machine-readable NTSE papers into structured, machine-readable PDFs.
- **High-quality dataset** that ensures accessibility and effective analysis.
- **Multiple data sources** are used to curate a comprehensive and reliable dataset.



EXAMPLES OF QUESTIONS

Cube and Dice Type Category Question



Question Figure

How many cubes are there in the 3D model Question Figure?

Embedded Figure Category Question



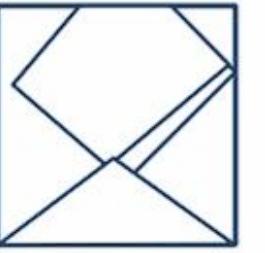
Question Figure



(1)



(2)



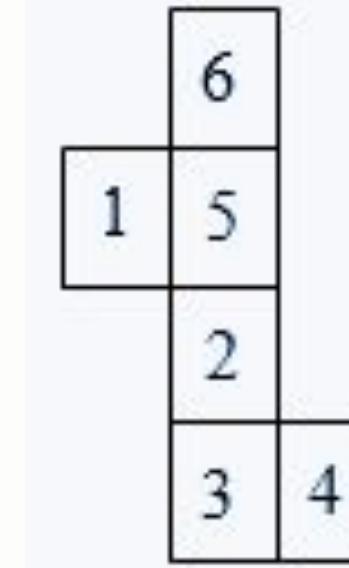
(3)



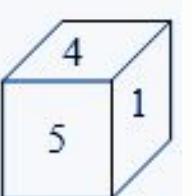
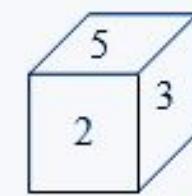
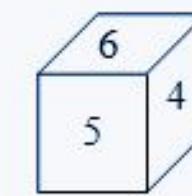
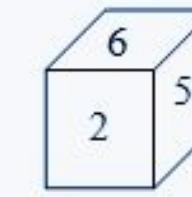
(4)

In which figure is the Question Figure embedded?

Paper Folding and Cutting Category Question



Question Figure



Which cube does the Question Figure yield on folding?

DATASET STATISTICS

- **Rich Multimodal Dataset:** Contains 2,728 multiple-choice questions, each paired with 4,642 images, forming comprehensive question–option–solution triplets.
- **Question Types:** We propose both **Text-only questions** as well as **Multimodal questions** incorporating images requiring deep logical analysis and reasoning

Question	Options	Solutions	No. of Samples
✗	✗	✗	1199
✗	✗	✓	381
✗	✓	✗	70
✗	✓	✓	18
✓	✗	✗	330
✓	✗	✓	126
✓	✓	✗	403
✓	✓	✓	201

A **tick (✓)** mark indicates that the question, option, or solution includes an image

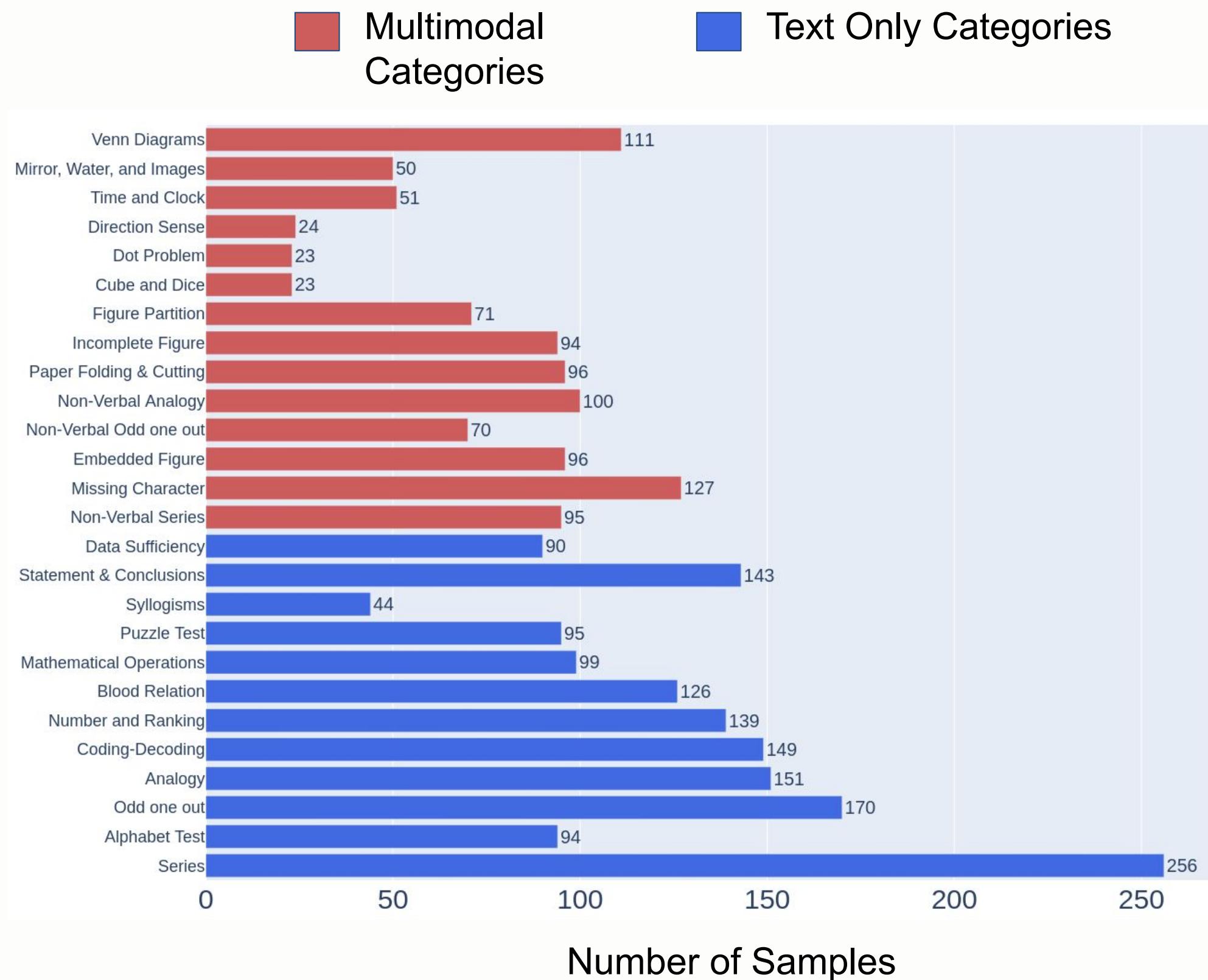
DATASET STATISTICS

- **Expert Categorization:** Data is meticulously

classified by human experts into **26 distinct categories**.

- **Cognitive Dimensions:** Introduces **8 cognitive**

dimensions to evaluate diverse aspects of multimodal reasoning.



- **Four context types** designed to handle different question types and input modalities
- **Models vary** in their ability to handle interleaving or multiple images
- **Ensures fair assessment** of model reasoning across different capabilities
- For **Text-only questions**, standard QA is employed.
- For **Multimodal questions**, 3 new types of strategies are proposed to cater to different models.

System Prompt

Question Text: In the number series given below, one number is missing.
\$ 12,15,27,42,69,111 \$,-

Option 1: 164 Option 2: 174 Option 3: 180 Option 4: 160

Answer format requested

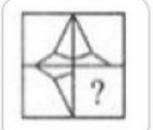
Standard QA: Used for Text-only questions

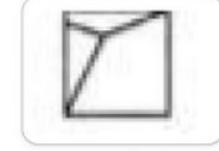
CONTEXT TYPES - Multimodal categories

- **Interleaved:** Images inserted into the text prompt, which preserves image **sequence and context**.
- **Standard VQA :** All images combined into one and labeled (e.g., Figure 1, 2, 3) to reference in text.
- **Image-Only:** Essentially a **snapshot of the question** as in PDF

System Prompt

Question Text: select a figure from amongst the four alternatives which when placed in the blank space of fig. (X) would complete the pattern.

Question Image: 

Option 1: 

Option 2: 

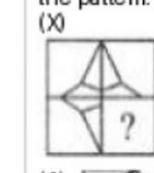
Option 3: 

Option 4: 

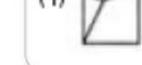
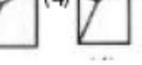
Answer format request

Interleaved

System Prompt

Question Image: 

select a figure from amongst the for alternatives which when placed in the blank space of fig. (X) would complete the pattern.
(X)

(1) 
(2) 
(3) 
(4) 

Answer format request

Image-Only

System Prompt

Question Image: 

Fig.1 
Fig.2 
Fig.3 
Fig.4 
Fig.5

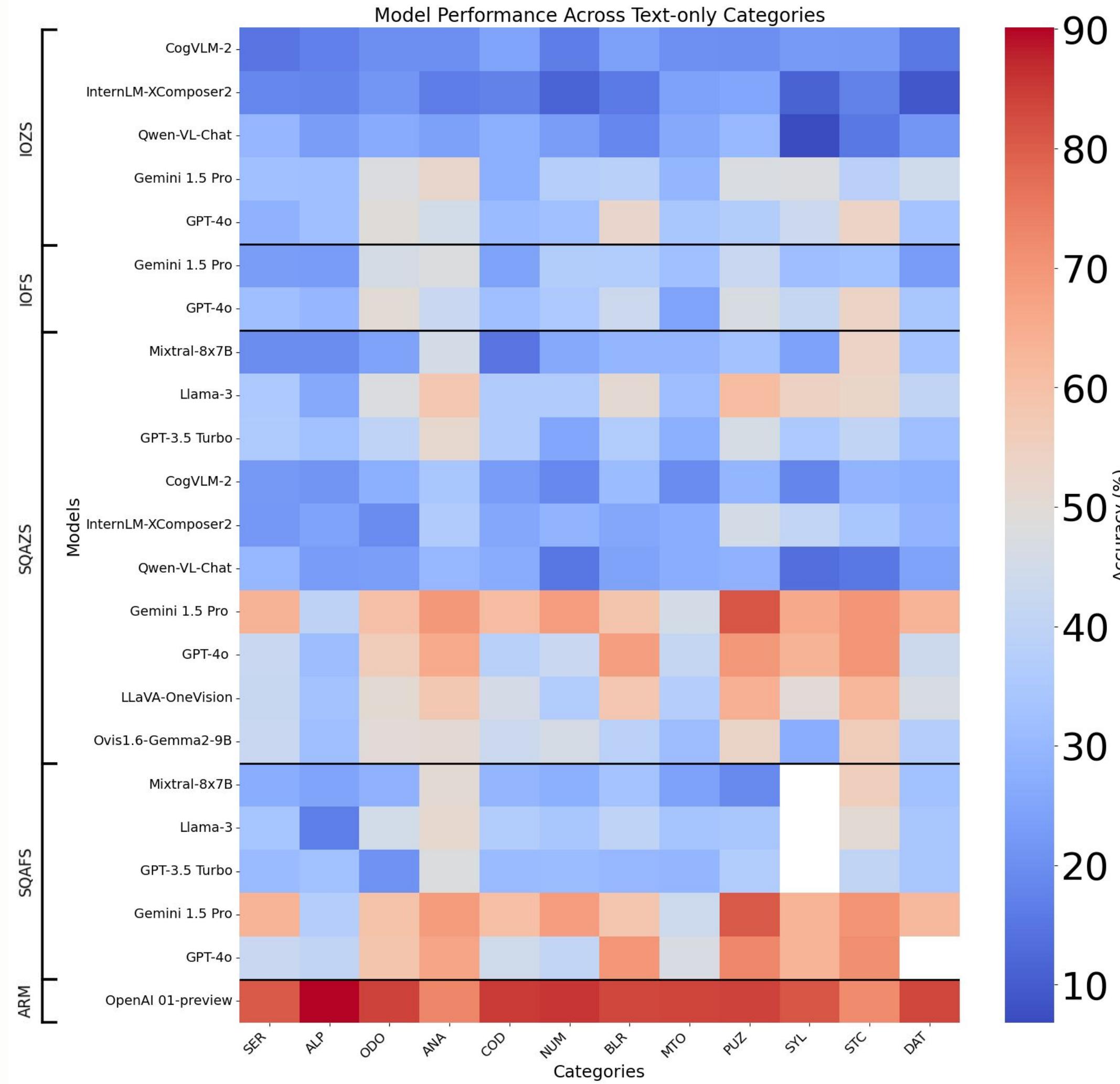
Question Text: select a figure from amongst the four alternatives which when placed in the blank space of fig. (X) would complete the pattern.
The image for question is as in Fig.1
Option 1: The image for option 1 is as in Fig.2
Option 2: The image for option 2 is as in Fig.3
Option 3: The image for option 3 is as in Fig.4
Option 4: The image for option 4 is as in Fig.5

Answer format request

Standard VQA

RESULTS - Text-only Categories Performance

IO: Image Only
SQA: Standard QA
ZS: Zero Shot
FS: Few Shot



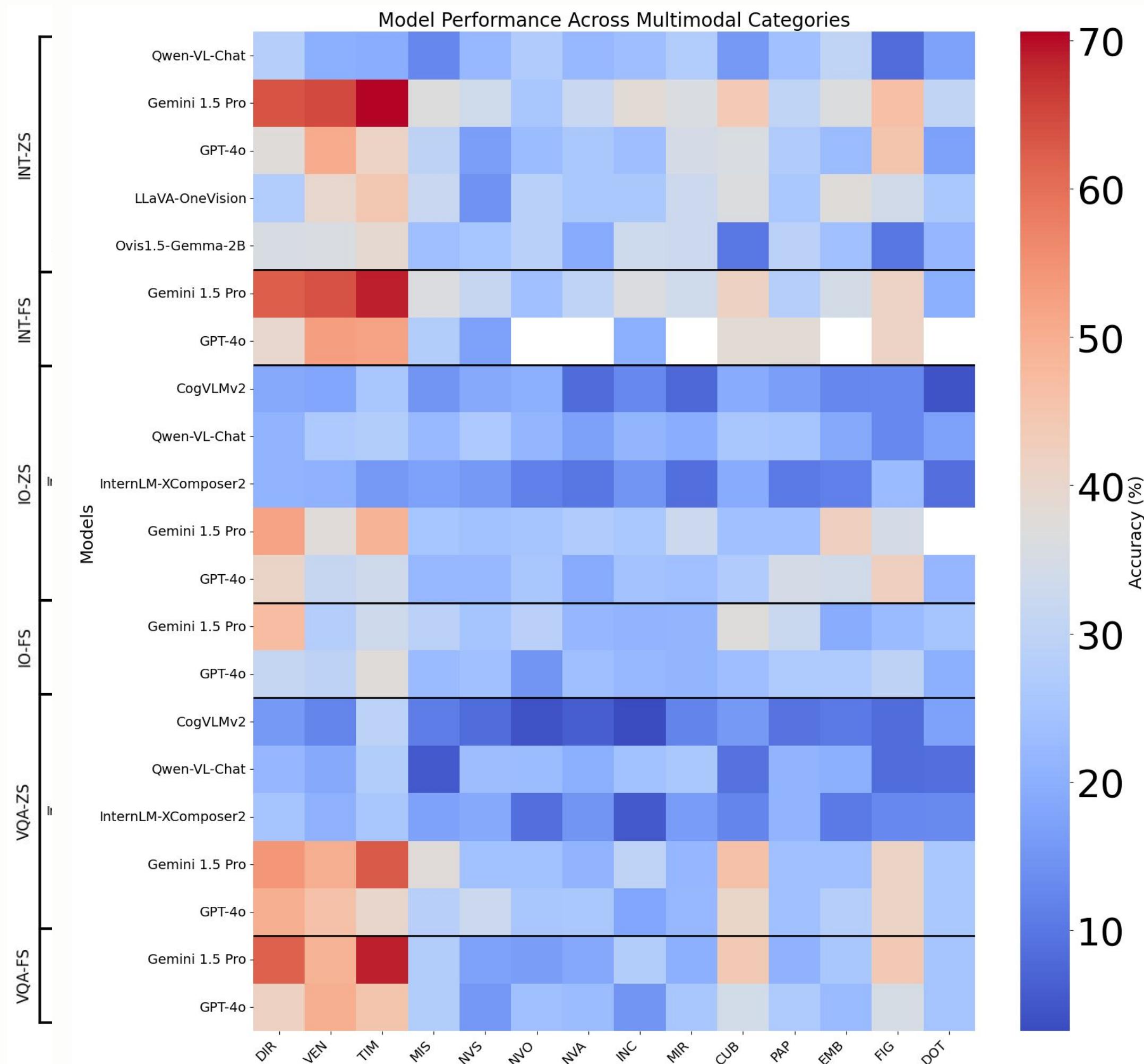
RESULTS -

Multimodal

Categories

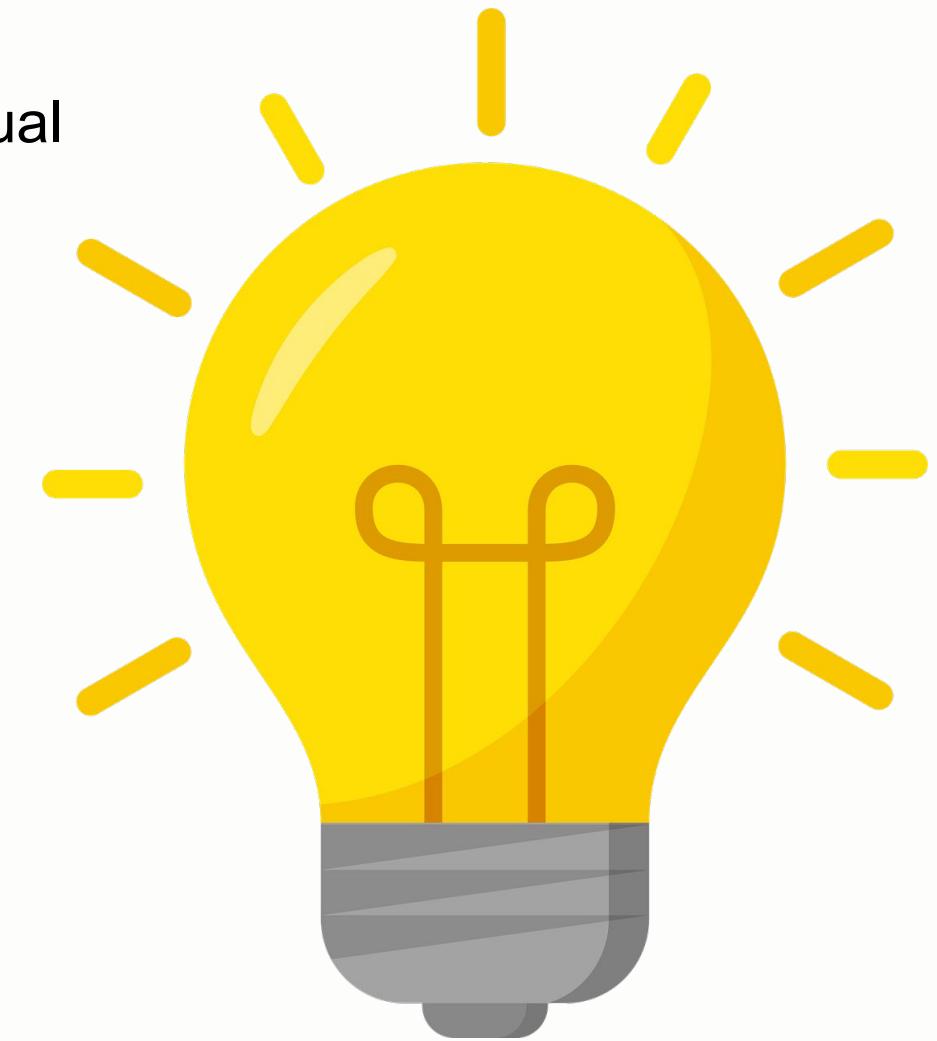
Performance

INT: Interleaving
IO: Image Only
VQA: Visual QA
ZS: Zero Shot
FS: Few Shot



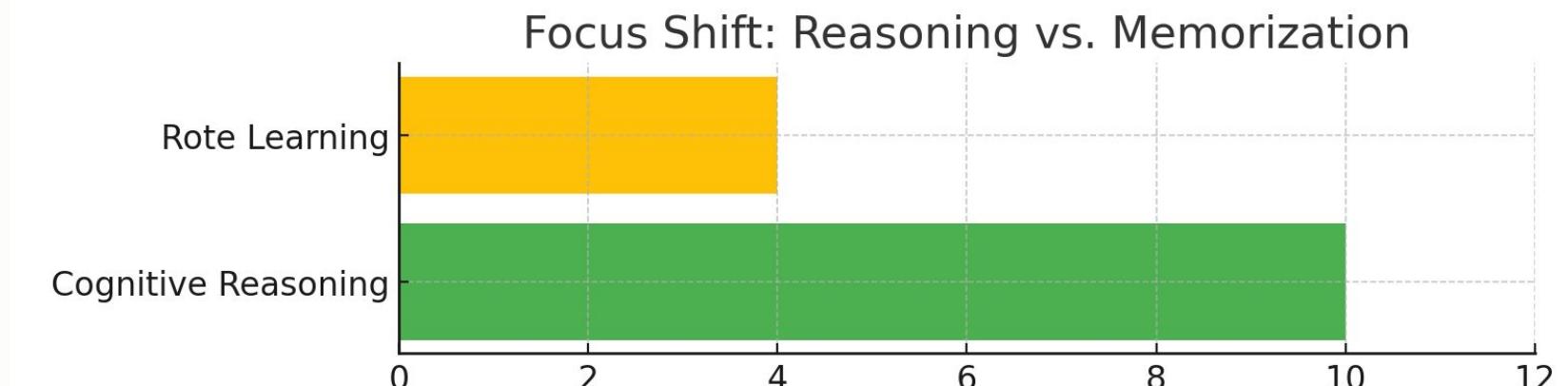
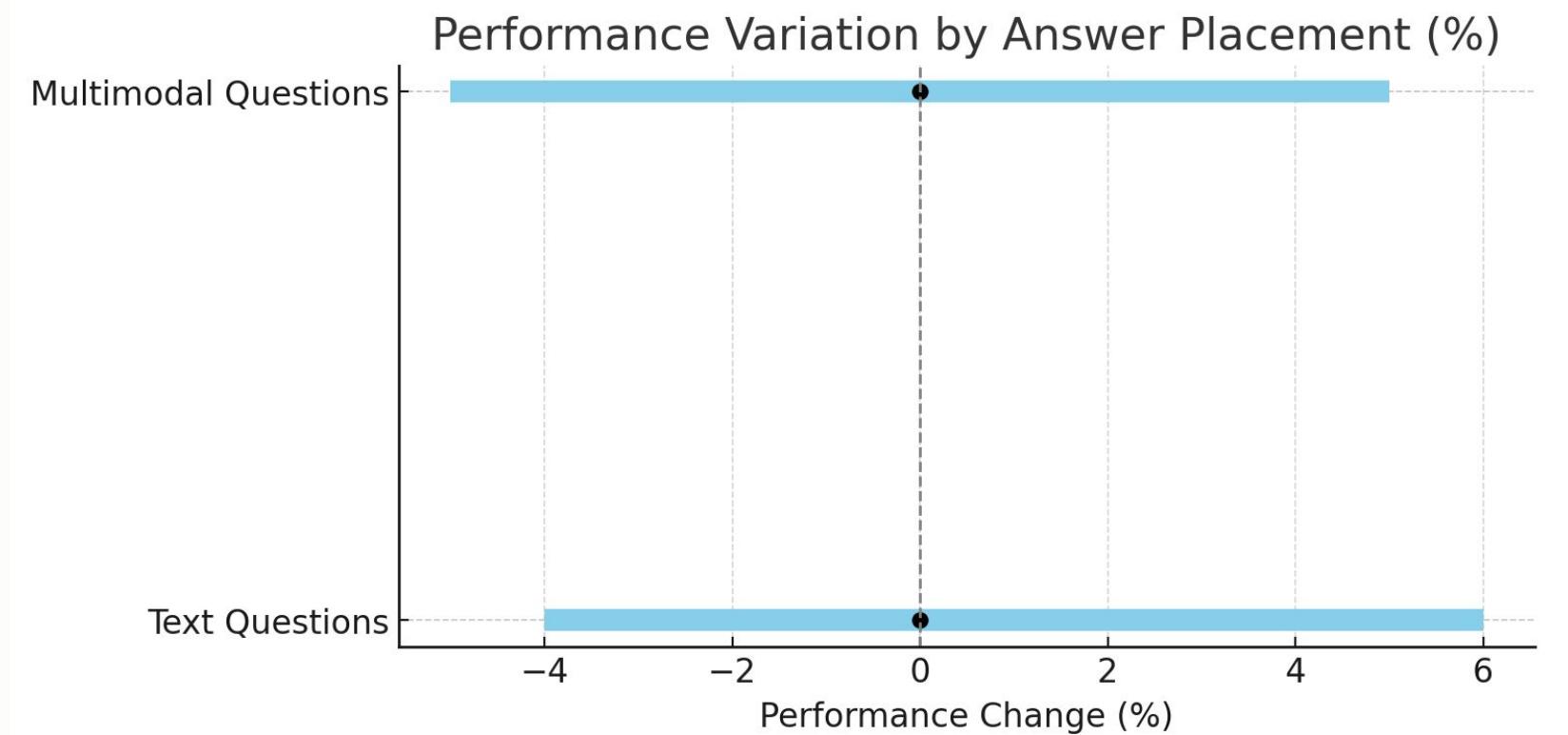
RESULTS—KEY INSIGHTS

- **Interleaving as a superior context mechanism**
 - **Showcased by ovis-1.6:** Outperforms larger open-source models due to better contextual representation.
- **Some categories** are relatively easier
 - Possibly due to high exposure in model pretraining datasets for **DIR, VEN, TIM.**
- **o1-preview's Text-Only Brilliance**
 - Showcases near-human performance on text-only tasks
 - Reinforces the importance of "**thinking before answering**" which is crucial in **NTSEBench**



OPTION ABALATION-BIAS EXPERIMENTATION

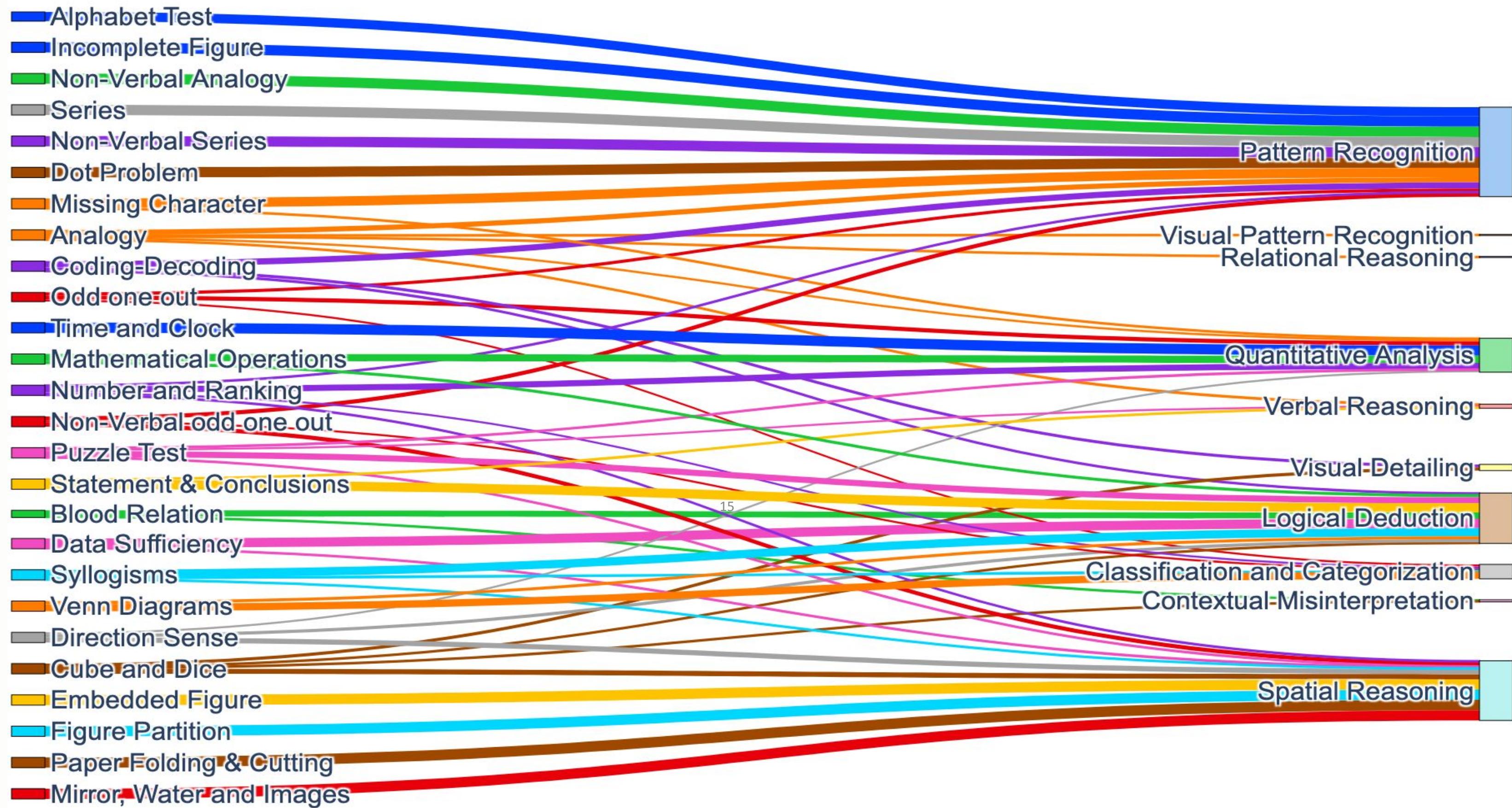
- **Robust analysis** accounting for cognitive reasoning rather than reliance on memorization.
- **Evaluated Gemini 1.5 Pro** to determine how answer placement affects model performance.
- **Observed performance shifts**
 - **-4% to +6%** for text-only questions.
 - **-5% to +5%** for multimodal questions.



EXTENSIVE ERROR ANALYSIS

- **Analysed 260 questions 10 per category** on Gemini 1.5 Pro, revealing reasoning patterns.
- **Categorized errors into the 8 cognitive dimensions proposed.**
- **VLMs struggle with logical deductions from limited visuals**, especially in pattern recognition, spatial manipulation, and shape recognition.
- Error distribution highlights **model strengths and scope for improvement**.

EXTENSIVE ERROR ANALYSIS



TAKEAWAYS

- **Dataset Expansion:**
 - **Data-Augmentation** techniques can be looked into to add more data for model training
 - **Hardness Categorization:** Deeper analysis to categorize hardness of questions can be done.
 - **Multilingual Capabilities:** Extend dataset to regional languages leveraging NTSE's multilingual reach.
- **Model & Method Insights:** Our framework allows for comprehensive evaluation across models and strategies, and can be extended to other datasets.
- **Challenging Benchmark:**
 - **Human accuracy (80%) > SOTA models (62% Text-only questions and 42% Multimodal questions)**
 - **NTSEBench proves itself to be a novel and important benchmark** that can improve models significantly and exposes model limitations in diverse categories.

DISCUSS

- We would be happy to discuss and address any questions.
- Please reach out to me at g.vatsal@alumni.iitg.ac.in; I would be happy to collaborate.

GITHUB



<https://github.com/NTSEBench/NTSEBench>

PAPER



<https://arxiv.org/abs/2407.10380>

WEBSITE



<https://ntsebench.github.io/>